Computer Methods and Programs in Biomedicine 251 (2024) 108213

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

NeuroDM: Decoding and visualizing human brain activity with EEG-guided diffusion model☆

Dongguan Qian [a], Hong Zeng [a], Wenjie Cheng [a], Yu Liu [a], Taha Bikki [a], Jianjiang Pan [b,*]

[a] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, Zhejiang, China
[b] School of Sciences, Hangzhou Dianzi University, Hangzhou, Zhejiang, China

ARTICLE INFO

Keywords:
Electroencephalography
Diffusion model
Feature extraction
Image generation

ABSTRACT

Background and Objective: Brain–Computer Interface (BCI) technology has recently been advancing rapidly, bringing significant hope for improving human health and quality of life. Decoding and visualizing visually evoked electroencephalography (EEG) signals into corresponding images plays a crucial role in the practical application of BCI technology. The recent emergence of diffusion models provides a good modeling basis for this work. However, the existing diffusion models still have great challenges in generating high-quality images from EEG, due to the low signal-to-noise ratio and strong randomness of EEG signals. The purpose of this study is to address the above-mentioned challenges by proposing a framework named NeuroDM that can decode human brain responses to visual stimuli from EEG-recorded brain activity.

Methods: In NeuroDM, an EEG-Visual-Transformer (EV-Transformer) is used to extract the visual-related features with high classification accuracy from EEG signals, then an EEG-Guided Diffusion Model (EG-DM) is employed to synthesize high-quality images from the EEG visual-related features.

Results: We conducted experiments on two EEG datasets (one is a forty-class dataset, and the other is a four-class dataset). In the task of EEG decoding, we achieved average accuracies of 99.80% and 92.07% on two datasets, respectively. In the task of EEG visualization, the Inception Score of the images generated by NeuroDM reached 15.04 and 8.67, respectively. All the above results outperform existing methods.

Conclusions: The experimental results on two EEG datasets demonstrate the effectiveness of the NeuroDM framework, achieving state-of-the-art performance in terms of classification accuracy and image quality. Furthermore, our NeuroDM exhibits strong generalization capabilities and the ability to generate diverse images.

1. Introduction

The human brain has long been regarded as an exceedingly intricate intelligent system, and researchers have devoted extensive efforts to its study [1,2]. Within this realm, decoding brain responses to visual stimuli has emerged as a trending research topic with broad application prospects. For example, it can aid disabled people in improving their daily communication and facilitate their rehabilitation training. Nevertheless, due to the remarkable complexity of the human brain system, decoding the human brain activity remains a great challenge.

The most widely employed brain imaging modality is electroencephalography (EEG), primarily due to its low cost and high temporal resolution. EEG is the objective response to neural activity in the brain [3], when an individual is faced with different visual stimuli, the neuronal behavioral pattern of the brain will be reflected in the EEG signals with the stimulus changes, that is, changes in brain activity patterns can be captured with EEG signals when individuals are exposed to different visual stimuli. Therefore, decoding visually evoked brain activity from EEG signals and visualizing them into corresponding images has attracted much attention. However, the existing methods still have challenges in generating high-quality images from EEG as well as strong robustness, mainly because: (1) EEG data is limited, contains a lot of noise, has significant individual difference. Therefore, it is difficult to learn meaningful visual-related information from EEG.

☆ Code available at: https://github.com/DongguanQian/NeuroDM.

* Corresponding author.
E-mail address: mathpan@hdu.edu.cn (J. Pan).

https://doi.org/10.1016/j.cmpb.2024.108213
Received 11 January 2024; Received in revised form 30 April 2024; Accepted 3 May 2024
Available online 7 May 2024

(2) There is significant difference in the data space of EEG and images, making it difficult for existing methods to directly map EEG to images.

In earlier years, researchers usually employed shallow methods to extract features from EEG signals [4–6], followed by feature selection and classification. Since the rise of deep learning, researchers began to employ deep learning methods for feature extraction and classification of EEG signals, such as long short-term memory (LSTM) [7], recurrent neural networks (RNN) [8] and convolutional neural network (CNN) [9]. The CNN-based methods can extract EEG information from different channels at the same time or from the same channel over time, using the spatial and temporal kernels respectively [10,11]. Some methods based on LSTM and RNN can capture the temporal information in EEG signals [12,13]. Some fusion methods use CNN to extract spatial features and then employ RNN to learn temporal information from these features [14,15]. However, the above-mentioned methods do not fully exploit the spatial–temporal information in the raw EEG signals and lack strong generalization capability.

The Transformer, employing multi-head attention [16], has recently been introduced into the BCI field. Some studies combine CNN with transformers for EEG classification tasks [17–19], showing great potential in capturing the inherent spatial and temporal dependencies within EEG signals. Inspired by the studies above, we introduce the ConvTransformer Block (CT Block) into EEG feature extraction to fully explore the spatial–temporal information.

Furthermore, image generation methods are crucial for decoding human brain activity as well. Generative adversarial networks (GANs) [20] can generate high-quality images, but they lack diversity and are prone to training collapse. Variational Auto-encoders (VAEs) [21] exhibit relatively stable training, yet they generate images of lower quality and struggle to learn complex data distributions.

Recently, diffusion models such as denoising diffusion probabilistic models (DDPM) [22,23], have achieved remarkable success in tasks related to image generation. Diffusion models consist of both a forward process and a reverse process. In the forward process, the input image is converted to Gaussian noise by progressively adding noise to it. Gaussian noise can be nearly perfectly inverted to the original image through step-by-step denoising in the reverse process. The work in [24] has demonstrated that diffusion models can beat GANs and VAEs in terms of image synthesis performance. Zeng et al. [25] proposed a framework based on the diffusion model, capable of reconstructing EEG into high-quality images. However, the model of the above method involves a large number of parameters, resulting in high computational cost. To address this issue, we design a lightweight module, Neuro Block, which can improve the performance of the diffusion model in generating images with lower computational cost.

Above all, we propose a framework NeuroDM based on the diffusion model, which could decode and visualize human brain responses to visual stimuli from EEG. The main contributions of this work are as follows:

1. We propose a novel EEG visual-related feature extraction module EV-Transformer, which could fully leverage the spatial–temporal information in EEG signals by introducing the ConvTransformer Block (CT Block).
2. We design a Neuro Block and incorporate it into the U-Net model [26], and then utilize an EG-DM algorithm to visualize EEG visual-related features as the corresponding images.
3. We design an EEG decoding and visualization framework NeuroDM and validate its effectiveness on two different EEG datasets (one is a forty-class dataset [27], and the other is a four-class dataset [28]).

The structure of this paper is as follows. In Section 2, we present an overview of existing methods for EEG decoding and visualization, along with insights from research on conditional diffusion models. Section 3 provides a detailed overview of our proposed method. Our experiments and results are detailed in Section 4, while Section 5 delves into a discussion of our proposed method and the experimental results. Finally, the conclusion and future work are presented in Section 6.

2. Related work

We divide the related literature into three sections. In Section 2.1, we review the existing methods for visual stimulus recognition based on EEG. In Section 2.2, we discuss the current state of research in EEG visualization field. Finally, we bring attention to the recent advancements in conditional diffusion models used for image generation in Section 2.3.

2.1. EEG decoding

Many deep learning methods have been employed for EEG decoding, *i.e.*, extracting EEG features and classifying them. At the same time, numerous EEG datasets suitable for visual object analysis have become available. Spampinato et al. [27] collected EEG data from 6 subjects with image stimuli from 40 ImageNet object classes, and employed methods based on RNN and CNN to enable EEG-based automated visual classification. Zeng et al. [28] collected an EEG dataset from 26 participants, stimulated by images from 4 ImageNet object classes. They proposed a bimodal semantic feature extraction method to capture common semantic information in EEG-image pairs. Du et al. [29] proposed a neural decoding method named BraVL that uses multimodal learning of brain-visual-linguistic features and constructed trimodal matching datasets. Zheng et al. [30] extracted features from EEG signals using LSTM and RNN methods with the Swish activation function. Simultaneously, they employed CNN and residual network methods for feature extraction from images. Subsequently, a regression method was utilized to map image features onto EEG features, achieving notably high classification accuracy on the 40-class dataset. Zeng et al. [25] proposed an EEG semantic feature extraction method named EVRNet, achieving a classification accuracy of 99.74% for 40 classes. The Transformer with multi-head attention [16], has recently been introduced into EEG decoding and has shown promising performance. Bagchi et al. [17] proposed a novel deep learning framework that combines the transformer and CNNs, providing state-of-the-art performance for five different visual stimulus classification tasks. Xie et al. [18] incorporated positional embedding modules into the transformer, enhancing the classification performance of EEG signals. Zeynali et al. [31] proposed a Transformer-based model to extract the temporal and spectral features of EEG signals.

2.2. EEG visualization

The primary objective of EEG visualization is to generate more realistic and high-definition images using the learned EEG visual-related features. The emergence of increasingly powerful generative models has provided more reliable options for EEG visualization. Zeng et al. [28] proposed a dual conditional autoencoder framework (DCAE), which can generate corresponding images by using multimodal fused features extracted from EEG and images. Zheng et al. [32] employed an enhanced spectral normalization generative adversarial network to conditionally generate images consistent with the visual stimulus categories using the learned EEG features. Khare et al. [33] introduced a conditional ProGAN pipeline that can easily regenerate the image by using EEG signals as input. Kumari et al. [34] incorporated the capsule network with GANs to reconstruct images using decoded features from EEG signals evoked by visual stimuli of the MNIST dataset. Zeng et al. [25] designed a reconstruction method based on DDPM, which can reconstruct the extracted EEG semantic features into images.

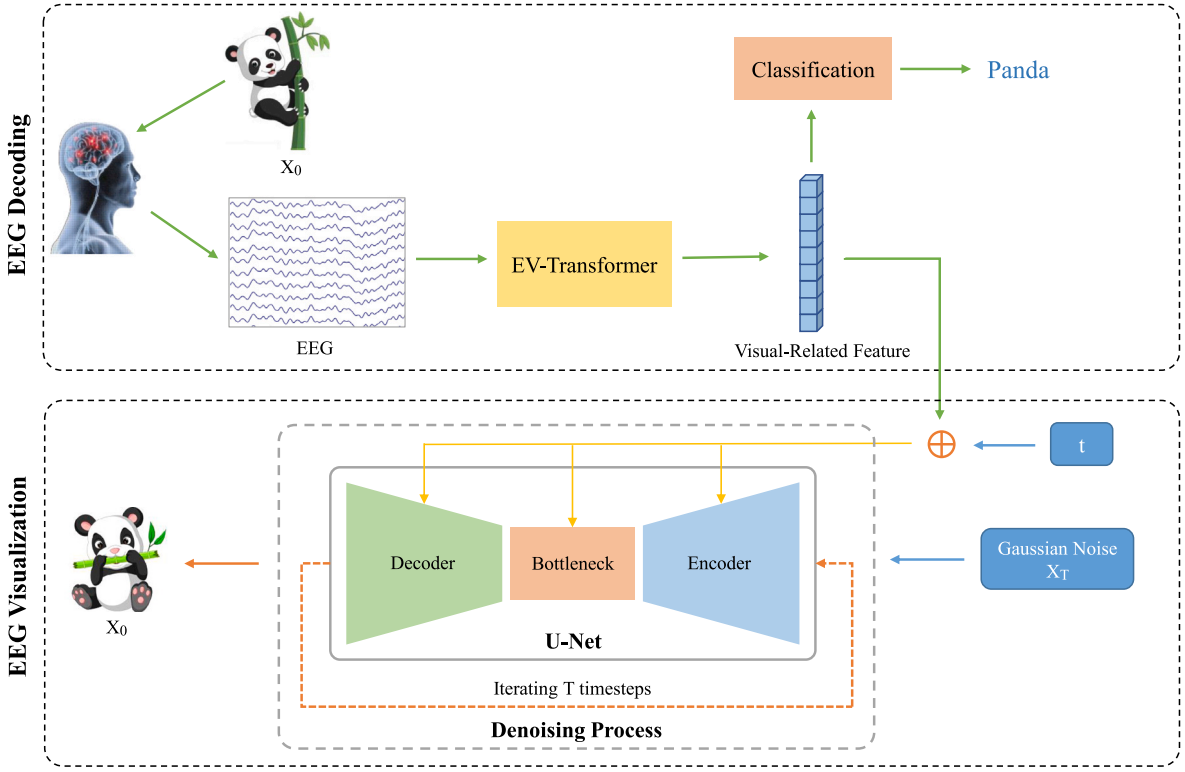

Fig. 1. The overall framework diagram of NeuroDM.

2.3. Conditional diffusion models

Diffusion models, as likelihood-based models, have been shown to achieve impressive results in image synthesis, particularly in conditional image synthesis, reaching the state-of-the-art level. Rombach et al. [35] presented latent diffusion models, which significantly enhance the training and sampling efficiency of denoising diffusion models while maintaining quality. By introducing cross-attention layers into the diffusion model, they enabled the model to handle general conditional inputs such as text and achieve high-resolution image synthesis. Dhariwal et al. [24] employed a classifier-guided technique, which allows the diffusion model to conditionally guide samples towards labels during the diffusion sampling process using classifier labels. Wang et al. [36] proposed a novel framework named Semantic Diffusion Model, which is based on DDPM and designed for semantic image synthesis. To further improve the quality of generated images and their semantic interpretability, they introduced a classifier-free guidance strategy during the sampling process. Kim et al. [37] employed another guiding strategy, Contrastive Language-Image Pretraining (CLIP) guidance, to guide the pretrained diffusion model for text-guided image manipulation. Nichol et al. [38] applied guided diffusion to text-conditional image synthesis, and compared the above-mentioned two guiding techniques. Ultimately, they found that the classifier-free guidance resulted in more realistic images. Inspired by this, we propose a novel guidance strategy, EEG guidance, to guide the diffusion model for generating images matching EEG.

3. Method

As shown in Fig. 1, our proposed NeuroDM framework consists of two parts, namely EEG Decoding and EEG Visualization. In the EEG Decoding section, the EV-Transformer decodes EEG into visual-related features, guided by the class labels of corresponding images. In the EEG Visualization section, we embed visual-related features into the timestep information, guiding the U-Net model for denoising, ultimately generating images corresponding to the respective class.

3.1. Preliminary of diffusion model

Diffusion probabilistic models [22] are latent variable models consisting of a forward process and a reverse process. In the forward process, the input image x_0 is converted to standard Gaussian noise by progressively adding Gaussian noise to it. The forward process can be expressed as:

$$q(x_t|x_0) = N(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)\boldsymbol{I}) \tag{1}$$

where t denotes the timestep of each noise addition, and $\boldsymbol{I}$ is the variance of image data x_0.

$$\alpha_t := \prod_{i=1}^{t}(1-\beta_i) \tag{2}$$

where β is a variance schedule. Therefore, by combining x_0 and Gaussian noise ϵ, we can obtain x_t:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon \tag{3}$$

Standard Gaussian noise can be nearly perfectly inverted to the original image through a step-by-step denoising in the reverse process. The reverse denoising process utilizes the U-Net architecture θ to effectively model the noise added at each timestep t:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \boldsymbol{\mu}_\theta(x_t,t), \boldsymbol{\Sigma}_\theta(x_t,t)) \tag{4}$$

The training objective of the conditional DDPM is to optimize the upper variational bound of the negative log likelihood. ϵ_θ is a model that predicts noise, and this predicted noise is used in the denoising process. And the mean $\boldsymbol{\mu}_\theta(x_t,t)$ can be calculated through a function of $\epsilon_\theta(x_t,t)$. Finally, a simplified training objective is obtained as follows:

$$L_{\text{simple}}(\theta) = \boldsymbol{E}_{t,x_0,\epsilon}[\|\epsilon - \boldsymbol{\epsilon}_\theta(x_t,t)\|] \tag{5}$$

where t is the timestep uniform between 1 and T. After training ϵ_θ, the following denoising process is employed for sampling:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_\theta(x_t,t)\right) + \sigma_t\epsilon \tag{6}$$

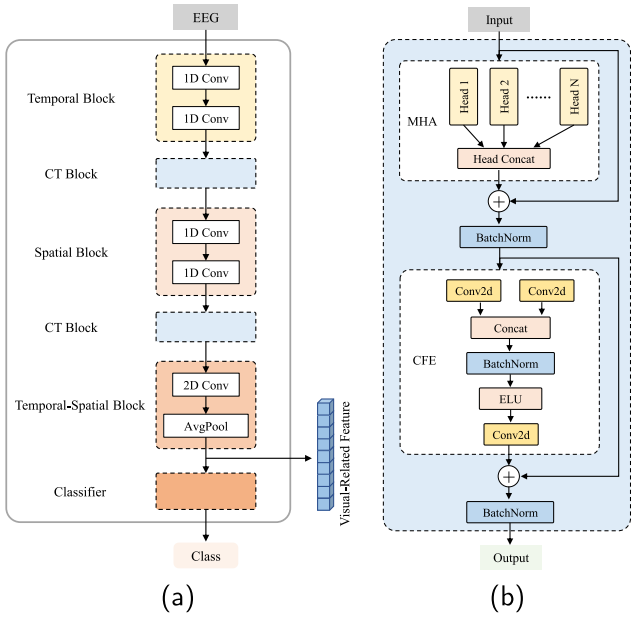

Fig. 2. (a) The structure of the EV-Transformer. (b) The structure of the CT Block.

where $\epsilon \sim N(0, \boldsymbol{I})$.

3.2. EEG decoding

In this part, we intend to extract visual-related feature from EEG sequence. Many previous studies [10–13] have focused solely on temporal and spatial information, lacking the ability to extract global information. To fully extract global features from the raw EEG signals, we design the EV-Transformer, whose structure is illustrated in Fig. 2(a). Within it, we introduce the CT Block, consisting of Multi-Head Attention (MHA) module and Convolutional Feature Expansion (CFE) module, as shown in Fig. 2(b).

The EV-Transformer takes EEG data as input and outputs visual-related features in size of $512 \times 1 \times 1$, trained using gradient descent by computing the cross-entropy loss between the predicted labels and the true labels. We illustrate the workflow of this section using the ImageNet-EEG-40 dataset as an example, as follows.

Firstly, the EEG data in size of $1 \times 128 \times 440$ is put into the Temporal Block, which consists of two convolutional layers with a kernel size of (5, 1) and a stride of (2, 1). A tensor with a size of $64 \times 128 \times 110$ is generated, and then it is fed into the first CT Block, maintaining the output size consistent. The obtained tensor is put into the Spatial Block, which includes two convolutional layers with a kernel size of (1, 5) and a stride of (1, 2). A tensor in size of $256 \times 32 \times 110$ is obtained and then put into the second CT Block with the same output size. After passing through the Temporal–Spatial Block consisting of a convolutional layer with a kernel size of (5, 5) and a stride of (2, 2) and an Average Pooling layer (AvgPool), the visual-related feature in size of $512 \times 1 \times 1$ is obtained. Finally, the visual-related feature is sent to the Classifier, which consists of a fully connected layer and a Softmax layer. BatchNorm normalization and ReLU activation are applied after each convolutional layer.

To fully extract global features from the EEG signals, we utilize the CT Block, which includes MHA module and CFE module, visualized in Fig. 2(b). The CT Block is crucial for capturing intricate patterns and relationships within the input tensor. After being processed by each self-attention head, the input tensor is concatenated along the channel axis to form a collection of multiple heads in the MHA module. Then the residual mapping is computed by adding the processed tensor to the original input tensor, followed by BatchNorm to get the input of the CFE module. In the CFE module, the input tensor is separately processed by two different convolutional layers, followed by channel concatenation. Next, BatchNorm and ELU activation are applied before the point-wise convolution. The final output of the CT Block is obtained after a residual mapping and BatchNorm.

3.3. EEG visualization

The EEG visualization part consists of the EG-DM algorithm and the U-Net model used for denoising. The EG-DM algorithm is employed to guide the U-Net model in the denoising process. Specifically, it utilizes visual-related feature extracted from EEG to guide the U-Net model in iteratively denoising Gaussian noise, ultimately generating images corresponding to the respective class, as illustrated in the lower part of Fig. 1.

3.3.1. EG-DM algorithm

The existing diffusion models [22,24] employ a shared U-Net [26] architecture θ for all t, inserting the information of t by using the sinusoidal position embedding used in the Transformer [16]. Meanwhile, we embed the EEG visual-related feature e into the information of t through a linear mapping. Combining with Eq. (5), we can obtain the objective loss function during training:

$$L(\theta) = \boldsymbol{E}_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(x_t, e, t)\|] \tag{7}$$

After training θ, we can iteratively denoise from Gaussian noise x_T to image x_0 during sampling. The denoising process is as follows:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, e, t)\right) + \sigma_t \epsilon \tag{8}$$

where $\epsilon \sim N(0, \boldsymbol{I})$.

3.3.2. U-Net model

Fig. 3 gives an overview of the U-Net model used for denoising in NeuroDM, which estimates the noise in the input noisy image. As depicted in Fig. 3, the U-Net model consists of three parts: Encoder, Bottleneck and Decoder. Firstly, the noisy image is fed into the Encoder, which encode the feature of the noisy image with 4 groups of Neuro Block, ResAttention and DownSample. Next, the Bottleneck further reinforces the encoded feature with two Neuro Blocks and a ResAttention. The reinforced feature is then fed into the Decoder for feature decoding with 4 groups of Neuro Block, ResAttention and UpSample. Meanwhile, the input feature of each layer in the Decoder is concatenated with the output feature from the corresponding layer in the Encoder. Finally, the decoded feature can generate denoised image after passing through the Output layer composed of StarReLU and NeuroBlock. To fully leverage the EEG visual-related feature e, the joint embedding of e and t is injected into each NeuroBlock.

ResAttention combines the characteristics of residual connection and attention mechanism, allowing it to capture crucial features of the input data while preserving the original information, as illustrated in Fig. 4. While DownSample reduces the size of the input by half along the spatial dimensions through the application of a 2D convolution with a kernel size of (4, 4) and a stride of (2, 2), UpSample performs the opposite operation by applying a 2D transposed convolution operator. To fully leverage semantic information to generate images with high-quality and strong semantic relevance, we design the Neuro Block to embed visual-related feature e into the denoising network. We show the detailed structure of the Neuro Block in Fig. 5, which consists of convolution, StarReLu and layer normalization. StarReLU [39] is a novel activation function that reduces the Floating Point Operations (FLOPs) of activation and achieves better performance compared with GELU, which can be expressed as

$$\mathrm{StarReLU}(x) = s \times (\mathrm{ReLU}(x))^2 + b \tag{9}$$

where s and b are scalars of scale and bias respectively, which are shared for all channels and can be set as constants or learnable parameters.

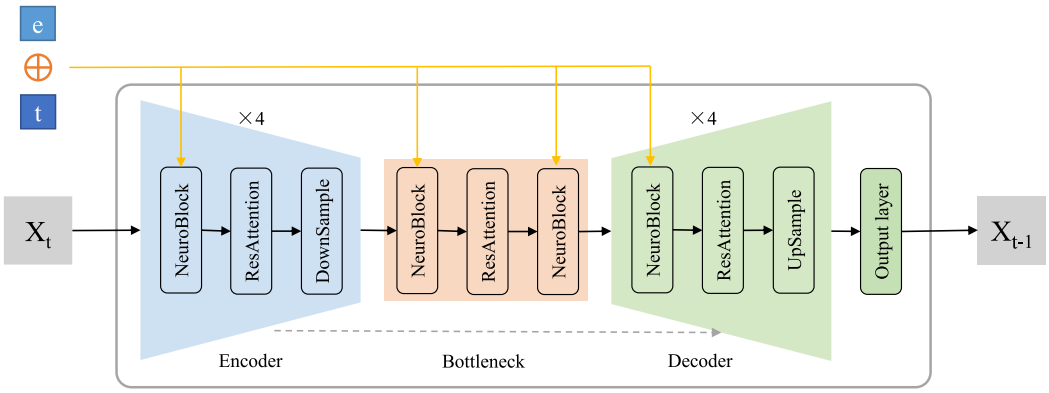

Fig. 3. The architecture of our U-Net model.

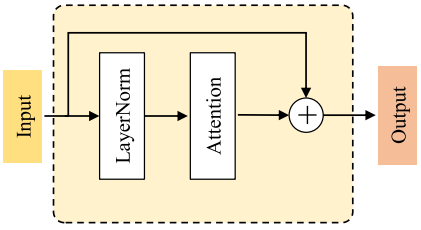

Fig. 4. The architecture of ResAttention.

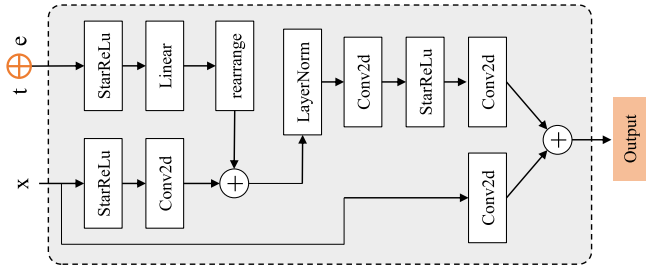

Fig. 5. The architecture of Neuro Block.

4. Experiments and results

4.1. Dataset

To evaluate our proposed NeuroDM and conduct comparative analysis with the existing methods, we adopt two EEG datasets both evoked by the image stimuli from ImageNet dataset, one is a publicly available forty-class dataset (ImageNet-EEG-40) [27], and the other is a four-class dataset that we collected (ImageNet-EEG-4) [28].

ImageNet-EEG-40 was collected from six subjects while they were viewing images from 40 classes of the ImageNet dataset, with each class comprising 50 images. The EEG data were recorded from 128 electrodes at a sampling frequency of 1000 Hz, and each image was presented for a duration of 500 ms. After removing a portion of low-quality samples, the number of EEG recordings in the dataset is 11,964. We selected data from 20 to 460 ms in each EEG recording. Zeng et al. [25] compared the classification performance across different frequency bands (1–70 Hz, 55–95 Hz, 14–70 Hz, 5–95 Hz) in this dataset and found that the classification accuracy in the frequency range of 1–70 Hz was significantly higher than others. Therefore, we applied a second-order bandpass Butterworth filter with a frequency range of 1–70 Hz.

ImageNet-EEG-4 was collected from 26 subjects with 32 electrodes and a sampling rate of 128 Hz. The stimulus images are selected from 4 classes of the ImageNet dataset, with each class containing 50 images. Each image was viewed by each subject 4 times, with each viewing lasting 400 ms. Since the frequency of 70 Hz exceeds the upper limit of the second-order bandpass Butterworth filter in this dataset, then we selected EEG data in the frequency range of 1–60 Hz for analysis.

4.2. Extraction and classification of EEG visual-related feature

4.2.1. Experimental settings

To comprehensively evaluate the performance of our proposed EV-Transformer in extracting visual-related features from EEG, we conducted experiments on two different datasets. We randomly divided all EEG recordings in each dataset into training, validation, and test sets with a ratio of 8:1:1. In our classification experiments, the evaluation metrics employed include accuracy and F1-Score. During the training process, we employed the Adam algorithm [40] for optimization with an initial learning rate of 0.001 and a batch size of 64. The total number of training epochs was set to 100 and the dimension of the extracted EEG visual-related features was configured to 512.

4.2.2. Performance evaluation

The results of our classification experiments, along with comparisons to some existing state-of-the-art methods, are illustrated in Table 1. As observed, our EV-Transformer achieved average accuracies of 99.80% and 92.07% on ImageNet-EEG-40 and ImageNet-EEG-4, respectively, achieving the highest classification accuracy compared to other methods. Especially on ImageNet-EEG-4, the performance of the EV-Transformer has shown a significant improvement compared to other existing methods. Fig. 6 presents the confusion matrix on ImageNet-EEG-4. We observed high classification accuracy for each class, but noticed a lower accuracy of 88.18% for flowers, due to the relatively diverse categories, colors, and shapes of flowers.

The curves of accuracy and loss on the training and validation sets during training are displayed in Fig. 7. It can be observed that the accuracy and loss reach a stable state within 100 epochs. This demonstrates that our model exhibits good convergence.

Additionally, we use Gradient-weighted Class Activation Mapping (GRAD-CAM) [41] to illustrate the activation levels of the EV-Trans former on the input EEG across various regions in the spatial–temporal plane. The averaged GRAD-CAMs for four common classes in both datasets are presented in Fig. 8. We can observe that most regions are activated in the GRAD-CAMs.

4.2.3. Ablation study

In order to demonstrate the performance improvement of EV-Transformer in extracting visual-related feature with the introduction of the CT Block, we conducted ablation experiments with different model structural designs. From Table 2, we can observe that the performance of EV-Transformer has improved to varying degrees with

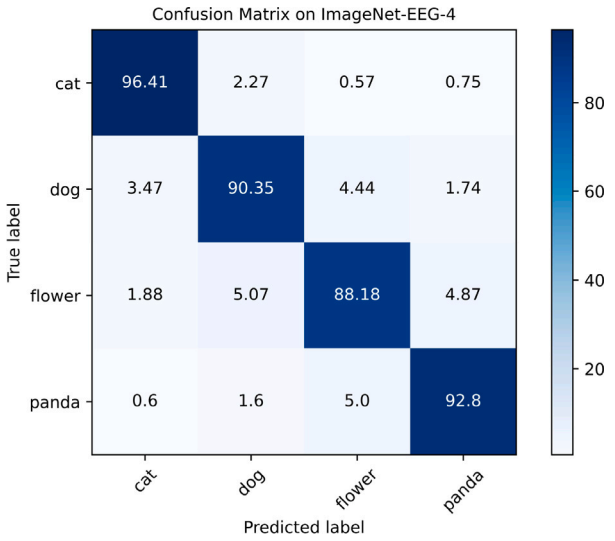

Fig. 6. The confusion matrix on ImageNet-EEG-4.

Table 1
Comparative experimental results.

Method	ImageNet-EEG-40		ImageNet-EEG-4	
	Acc (%)	F1-Score	Acc (%)	F1-Score
LSTM [42]	83.36	0.836	47.25	0.475
EEGNet [43]	88.13	0.881	48.65	0.486
SyncNet [44]	83.45	0.837	48.91	0.491
A-Bi-LSTM [45]	94.15	0.946	67.05	0.673
EEG-ChannelNet [46]	98.35	0.987	71.40	0.713
EV-Net [28]	98.98	0.988	80.42	0.807
EVRNet [25]	99.74	0.998	84.76	0.851
EV-Transformer	**99.80**	**0.998**	**92.07**	**0.919**

Table 2
Ablation experimental results.

Architecture	ImageNet-EEG-40		ImageNet-EEG-4	
	Acc (%)	F1-Score	Acc (%)	F1-Score
S+T+TS	92.98	0.927	85.87	0.859
T+S+TS	97.07	0.971	88.70	0.887
T+S+CT+TS	99.00	0.990	90.67	0.907
T+CT+S+TS	98.91	0.989	91.35	0.913
T+CT+S+CT+TS	99.80	0.998	92.07	0.919

*T: Temporal Block, S: Spatial Block,
TS: Temporal–Spatial Block, CT: CT Block.

the addition of CT Block. The results of the ablation study demonstrate the effectiveness of the CT Block in extracting global features from EEG signals.

4.3. Image generation using EEG visual-related feature

4.3.1. Experimental settings

During the image generation phase, we utilize the visual-related features extracted from EEG to generate images with the same class semantics. The training set and test set correspond to the data split in the previous stage, and within the training set, all visual-related features of the same class were averaged. To reduce computational cost, the resolution of all images in the dataset was adjusted to 64 × 64, and the resolution for generating images was also set accordingly. We adopted different timestep T and two variance schedules: Cosine schedule and Linear schedule. NeuroDM was trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 64. The total number of training epochs was set to 1000 for ImageNet-EEG-40 and 500 for ImageNet-EEG-4, respectively. Additionally, the Inception Score [47] is employed to assess the quality of the generated images, considering both clarity and diversity.

Table 3
Inception Score on ImageNet-EEG-40 and ImageNet-EEG-4.

Method	Inception score	
	ImageNet-EEG-40	ImageNet-EEG-4
BCVAE [48]	1.62	2.02
DCVAE [49]	1.98	1.92
DM-RE2I [25]	12.55	8.16
Our NeuroDM	15.04	8.67

Table 4
Inception Score on ImageNet-EEG-40.

Method	Inception score
DCGAN [50]	5.07
SNGAN [32]	5.53
DCLS-GAN [51]	6.64
DM-RE2I [25]	7.46
Our NeuroDM	15.89

Table 5
The FLOPs and params number of DM-RE2I and NeuroDM.

Method	FLOPs	Params
DM-RE2I [25]	8.43G	56.64M
Our NeuroDM	4.94G	41.47M

* $1M = 10^6, 1G = 10^9$.

Table 6
Inception Score on different activation function.

Activation function	Inception score	
	ImageNet-EEG-40	ImageNet-EEG-4
GELU	11.44	7.29
StarReLU	15.04	8.87

4.3.2. Performance evaluation

In order to better validate the performance of our NeuroDM, we employed two comparison methods when contrasting with some existing state-of-the-art methods. One method is generating an image for each EEG visual-related feature in the testset, with the results shown in Table 3. Another is generating 1250 images for each visual class (totaling 50,000 images), commonly used in GAN-based methods, with the results shown in Table 4. From the above results, we can observe that the quality of the generated images by Our NeuroDM is significantly superior to GAN-based methods, VAE-based methods, and also outperforms DM-RE2I.

Partial high-quality sampling results for each class on ImageNet-EEG-40 and ImageNet-EEG-4 are presented in Figs. 9 and 10, which include images seen by the subjects and generated images. From the figures, it can be observed that the images generated by our NeuroDM exhibit not only high resolution but also high semantic similarity to the original images. At the same time, the ability to generate high-quality images for each class indicates that our method possesses excellent diversity and robustness.

We also compared the Floating Point Operations (FLOPs) and the number of parameters on DM-RE2I and NeuroDM. The results in Table 5 show that Our NeuroDM has higher computational efficiency. The above results indicate that our model achieves superior performance with lower computational cost, making it a promising choice for deployment in resource-constrained environments.

In addition, we compared the impact of two different activation functions, GELU and StarReLU, on the quality of generated images. The results in Table 6 demonstrate that introducing StarReLU achieves better performance. Finally, we explored the impact of different variance schedules and timesteps on the quality of generated images. In theory,

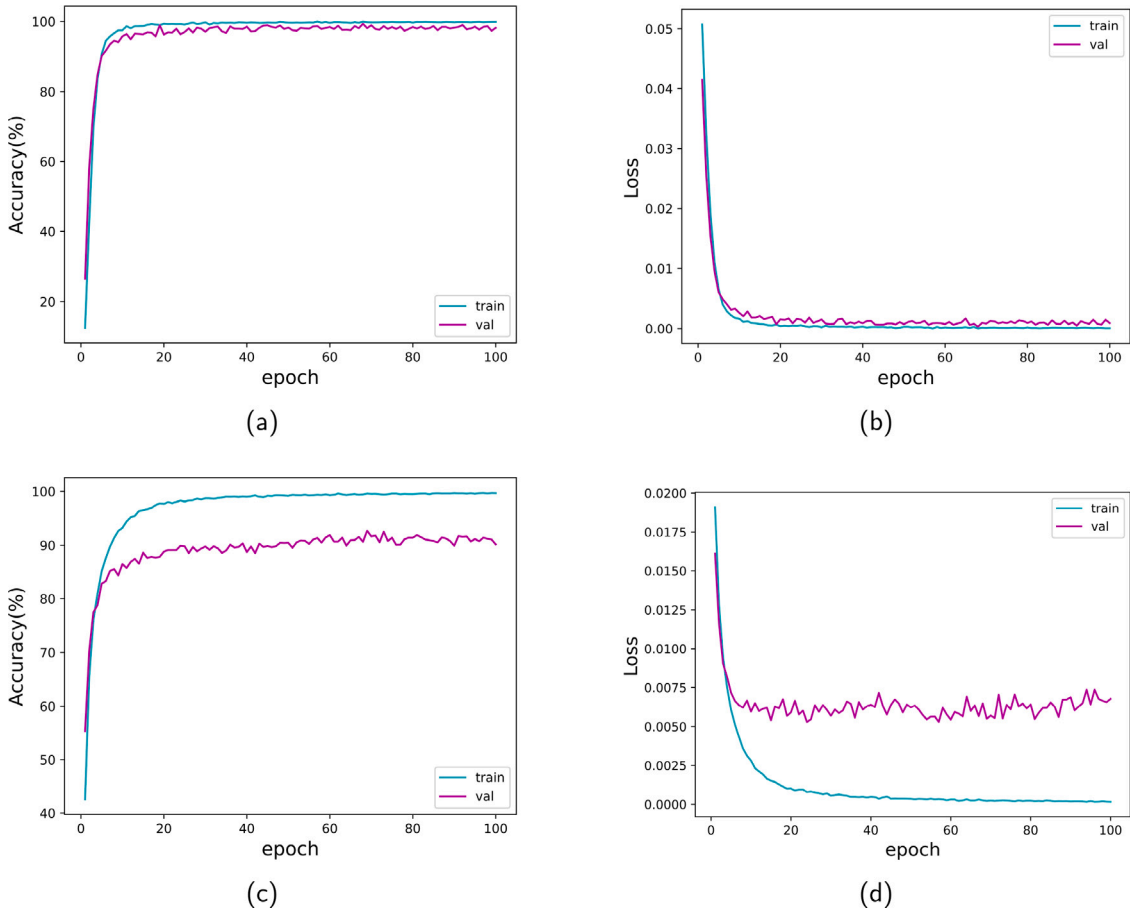

Fig. 7. Accuracy and loss on the training and validation sets during training of EV-Transformer. (a) The accuracy curve on ImageNet-EEG-40. (b) The loss curve on ImageNet-EEG-40. (c) The accuracy curve on ImageNet-EEG-4. (d) The loss curve on ImageNet-EEG-4.

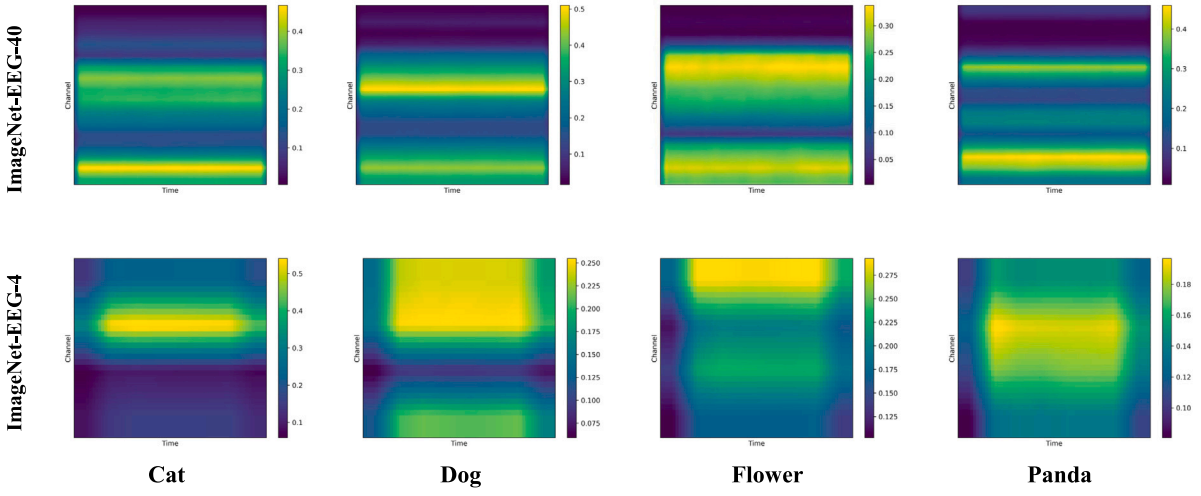

Fig. 8. The GRAD-CAM averaged for four common classes on ImageNet-EEG-40 and ImageNet-EEG-4. The horizontal axis represents time, with direction from left to right, while the vertical axis represents channels, indicating spatial dimensions, with direction from top to bottom.

the larger the value of the timestep T, the better the model is expected to be trained. However, in our experiments, we observed that setting T to a larger value does not necessarily improve performance. Instead, it would result in increased time consumption. The results in Table 7 illustrate that the impact of different configurations on performance is substantial, and the optimal configuration varies across different datasets.

5. Discussion

To visualize brain activity recorded in EEG as corresponding images, we first encode EEG signals into low-dimensional meaningful visual-related features using EV-Transformer with high classification accuracy. In this stage, we introduce the CT Block, which utilizes

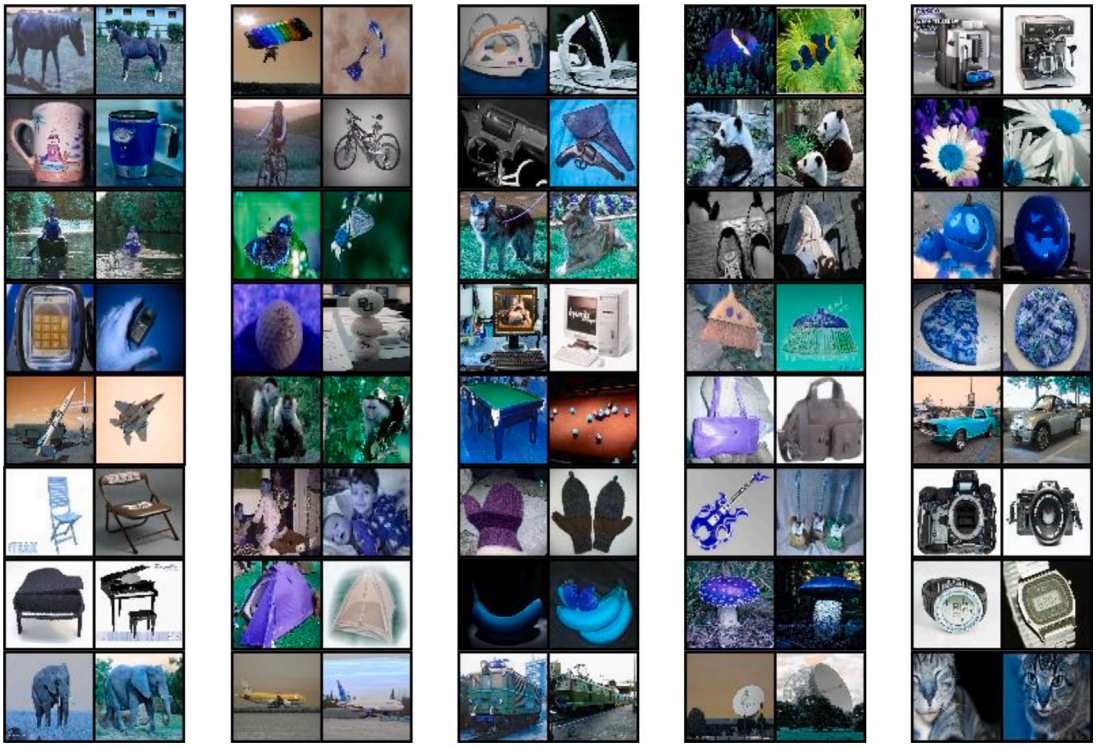

Fig. 9. Partial high-quality sampling results on ImageNet-EEG-40. There are 40 pairs of images, each representing a different class. In each pair, the images on the left are seen by the subjects, and on the right are generated by NeuroDM based on EEG.

Table 7
Inception Score on different configurations.

Schedule	T	Inception score	
		ImageNet-EEG-40	ImageNet-EEG-4
Linear	10	13.71	8.09
Linear	30	14.08	8.48
Linear	50	15.04	8.45
Linear	80	13.58	8.68
Cosine	10	10.41	8.19
Cosine	30	12.06	7.84
Cosine	50	13.71	8.87
Cosine	80	13.54	8.29

multi-head attention, to better extract global features from EEG, including both temporal and spatial information. The results in Table 2 demonstrate the effectiveness of the CT Block. From Fig. 8, we observe that most regions are activated in the GRAD-CAMs, indicating the effectiveness of our feature extraction method.

In the EEG visualization stage, we employ the EEG-Guided Diffusion Model, namely NeuroDM, given the powerful performance demonstrated by diffusion models in recent image generation tasks. We utilize visual-related feature extracted from EEG to guide the U-Net model in denoising process to generate images corresponding to the respective class. To fully leverage semantic information, we design the Neuro Block to embed visual-related feature into the denoising network. The experimental results show that the image quality generated by our NeuroDM is significantly superior to GAN-based methods and VAE-based methods. Compared to DM-RE2I, NeuroDM achieves superior performance at a lower computational cost. The EEG visualization results in Figs. 9 and 10 demonstrate that our method can generate diverse and high-quality images.

We observed that the performance of our method in ImageNet-EEG-40 was better than that of ImageNet-EEG-4, both in terms of classification accuracy and image generation quality. This is mainly because the EEG data in ImageNet-EEG-40 has more channels and higher temporal resolution, containing more information about the visual stimuli.

In addition, we adopt a novel activation function StarReLU and demonstrate it could achieve better performance in image generation, as shown in Table 6. Finally, we compared the impact of different variance schedules and timesteps on performance, with the results presented in Table 7. We found that the impact of different configurations on performance is substantial, and the optimal configuration varies across different datasets. This indicates that our method has the potential to generalize to more datasets by simply adjusting the configurations.

6. Conclusion

This paper proposes a novel method, NeuroDM, for decoding and visualizing human brain activity from EEG signals. It consists of two stages: EEG decoding and EEG visualization. In the EEG decoding stage, we transform EEG signals into low-dimensional visual-related features using EV-Transformer. In the EEG visualization stage, we employ NeuroDM to reconstruct EEG visual-related features into corresponding images. The experimental results indicate that our proposed method

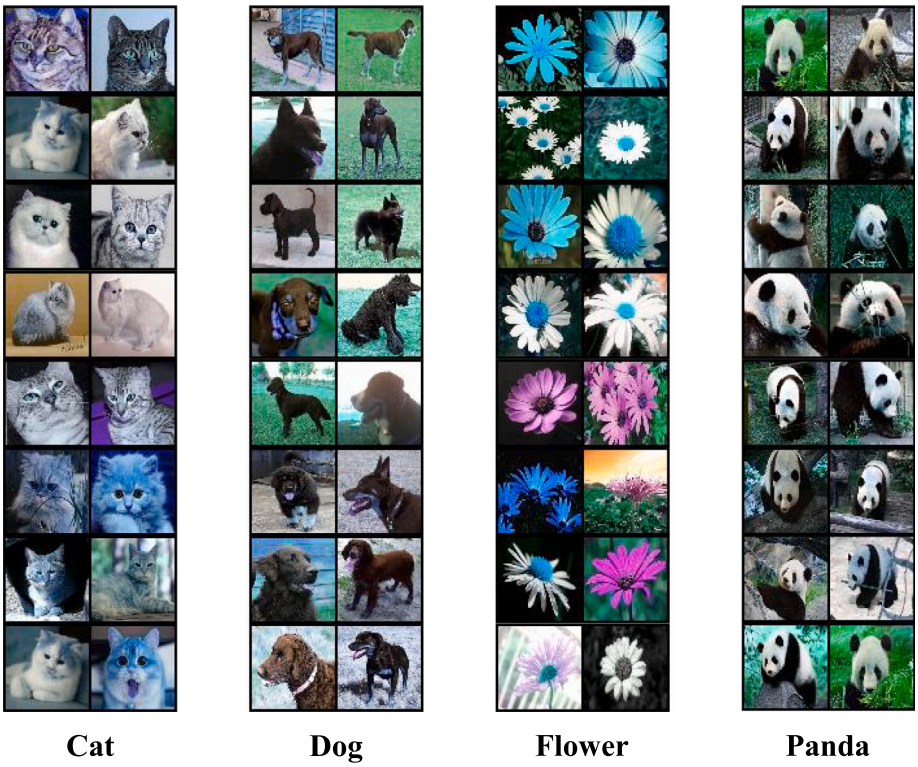

Fig. 10. Some high-quality sampling results on ImageNet-EEG-4. Here are 4 columns of images, each representing a different class. In each column, the images on the left are those seen by the subjects, while the images on the right are generated by NeuroDM based on EEG.

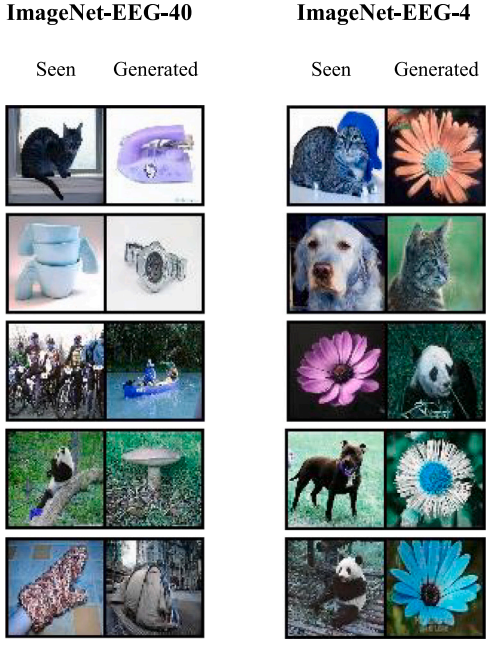

Fig. 11. A few poor sampling results on ImageNet-EEG-40 and ImageNet-EEG-4.

achieves higher classification accuracy and superior image generation quality compared to existing methods. Furthermore, our NeuroDM exhibits strong generalization capabilities and the ability to generate diverse images. Currently, we have only utilized class-level limited information from EEG. Fig. 11 shows some poor sampling results, where the generated images do not match the classes of the images seen by the subjects. In the future, we can explore how to extract more fine-grained information from EEG for decoding and visualizing human brain activity, such as object colors, shapes, and background information.

CRediT authorship contribution statement

Dongguan Qian: Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Hong Zeng:** Writing – review & editing, Supervision, Resources, Funding acquisition, Formal analysis, Conceptualization. **Wenjie Cheng:** Validation, Investigation, Data curation. **Yu Liu:** Validation, Investigation, Data curation. **Taha Bikki:** Validation, Investigation, Data curation. **Jianjiang Pan:** Supervision, Project administration, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by the National Key Research and Development Program of China under grant No. 2022YFE0199300, the National Science Foundation of China under No. 62076083, Zhejiang Provincial Natural Science Foundation of China under Grant No. ZCLZ24F0301, the “Leading Goose” R&D Program of Zhejiang, China with grant No. 2023C03026.

References

[1] Koel Das, Barry Giesbrecht, Miguel P. Eckstein, Predicting variations of perceptual performance across individuals from neural activity using pattern classifiers, Neuroimage 51 (4) (2010) 1425–1437.

[2] Thomas A. Carlson, Hinze Hogendoorn, Ryota Kanai, Juraj Mesik, Jeremy Turret, High temporal resolution decoding of object position and category, J. Vis. 11 (10) (2011) 9.

[3] Petra Ritter, Arno Villringer, Simultaneous EEG–fMRI, Neurosci. Biobehav. Rev. 30 (6) (2006) 823–838.

[4] Sara Jafakesh, Fatemeh Zareayan Jahromy, Mohammad Reza Daliri, Decoding of object categories from brain signals using cross frequency coupling methods, Biomed. Signal Process. Control 27 (2016) 60–67.

[5] Mitra Taghizadeh-Sarabi, Mohammad Reza Daliri, Kavous Salehzadeh Niksirat, Decoding objects of basic categories from electroencephalographic signals using wavelet transform and support vector machines, Brain Topogr. 28 (2015) 33–46.

[6] Taban Fami Tafreshi, Mohammad Reza Daliri, Mahrad Ghodousi, Functional and effective connectivity based features of EEG signals for object recognition, Cogn. Neurodyn. 13 (2019) 555–566.

[7] Salma Alhagry, Aly Aly Fahmy, Reda A. El-Khoribi, Emotion recognition based on EEG using LSTM recurrent neural network, Int. J. Adv. Comput. Sci. Appl. 8 (10) (2017).

[8] Pouya Bashivan, Irina Rish, Mohammed Yeasin, Noel Codella, Learning representations from EEG with deep recurrent-convolutional neural networks, 2015, arXiv preprint arXiv:1511.06448.

[9] Manjunath Jogin, M.S. Madhulika, G.D. Divya, R.K. Meghana, S. Apoorva, et al., Feature extraction using convolution neural networks (CNN) and deep learning, in: 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology, RTEICT, IEEE, 2018, pp. 2319–2323.

[10] Hauke Dose, Jakob S. Møller, Helle K. Iversen, Sadasivan Puthusserypady, An end-to-end deep learning approach to MI-EEG signal classification for BCIs, Expert Syst. Appl. 114 (2018) 532–542.

[11] Xiaying Wang, Michael Hersche, Batuhan Tömekce, Burak Kaya, Michele Magno, Luca Benini, An accurate eegnet-based motor-imagery brain–computer interface for low-power edge computing, in: 2020 IEEE International Symposium on Medical Measurements and Applications, MeMeA, IEEE, 2020, pp. 1–6.

[12] Ping Wang, Aimin Jiang, Xiaofeng Liu, Jing Shang, Li Zhang, LSTM-based EEG classification in motor imagery tasks, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (11) (2018) 2086–2095.

[13] John Thomas, Tomasz Maszczyk, Nishant Sinha, Tilmann Kluge, Justin Dauwels, Deep learning-based classification for brain-computer interfaces, in: 2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC, IEEE, 2017, pp. 234–239.

[14] Weizheng Qiao, Xiaojun Bi, Deep spatial-temporal neural network for classification of EEG-based motor imagery, in: Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, 2019, pp. 265–272.

[15] Xinjie Shi, Tianqi Wang, Lan Wang, Hanjun Liu, Nan Yan, Hybrid convolutional recurrent neural networks outperform CNN and RNN in task-state EEG detection for parkinson's disease, in: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC, IEEE, 2019, pp. 939–944.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[17] Subhranil Bagchi, Deepti R. Bathula, EEG-ConvTransformer for single-trial EEG-based visual stimulus classification, Pattern Recognit. 129 (2022) 108757.

[18] Jin Xie, Jie Zhang, Jiayao Sun, Zheng Ma, Liuni Qin, Guanglin Li, Huihui Zhou, Yang Zhan, A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification, IEEE Trans. Neural Syst. Rehabil. Eng. 30 (2022) 2126–2136.

[19] Yonghao Song, Qingqing Zheng, Bingchuan Liu, Xiaorong Gao, EEG conformer: Convolutional transformer for EEG decoding and visualization, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2022) 710–719.

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, Timo Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8110–8119.

[21] Ali Razavi, Aaron Van den Oord, Oriol Vinyals, Generating diverse high-fidelity images with vq-vae-2, Adv. Neural Inf. Process. Syst. 32 (2019).

[22] Jonathan Ho, Ajay Jain, Pieter Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.

[23] Alexander Quinn Nichol, Prafulla Dhariwal, Improved denoising diffusion probabilistic models, in: International Conference on Machine Learning, PMLR, 2021, pp. 8162–8171.

[24] Prafulla Dhariwal, Alexander Nichol, Diffusion models beat gans on image synthesis, Adv. Neural Inf. Process. Syst. 34 (2021) 8780–8794.

[25] Hong Zeng, Nianzhang Xia, Dongguan Qian, Motonobu Hattori, Chu Wang, Wanzeng Kong, DM-RE2I: A framework based on diffusion model for the reconstruction from EEG to image, Biomed. Signal Process. Control 86 (2023) 105125.

[26] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[27] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, Mubarak Shah, Deep learning human mind for automated visual classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6809–6817.

[28] Hong Zeng, Nianzhang Xia, Ming Tao, Deng Pan, Haohao Zheng, Chu Wang, Feifan Xu, Wael Zakaria, Guojun Dai, DCAE: A dual conditional autoencoder framework for the reconstruction from EEG into image, Biomed. Signal Process. Control 81 (2023) 104440.

[29] Changde Du, Kaicheng Fu, Jinpeng Li, Huiguang He, Decoding visual neural representations by multimodal learning of brain-visual-linguistic features, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[30] Xiao Zheng, Wanzhong Chen, Yang You, Yun Jiang, Mingyang Li, Tao Zhang, Ensemble deep learning for automated visual classification using EEG signals, Pattern Recognit. 102 (2020) 107147.

[31] Mahsa Zeynali, Hadi Seyedarabi, Reza Afrouzian, Classification of EEG signals using Transformer based deep learning and ensemble models, Biomed. Signal Process. Control.

[32] Xiao Zheng, Wanzhong Chen, Mingyang Li, Tao Zhang, Yang You, Yun Jiang, Decoding human brain activity with deep learning, Biomed. Signal Process. Control 56 (2020) 101730.

[33] Sanchita Khare, Rajiv Nayan Choubey, Loveleen Amar, Venkanna Udutalapalli, NeuroVision: perceived image regeneration using cProGAN, Neural Comput. Appl. 34 (8) (2022) 5979–5991.

[34] Nandini Kumari, Shamama Anwar, Vandana Bhattacharjee, Sudip Kumar Sahana, Visually evoked brain signals guided image regeneration using GAN variants, Multimedia Tools Appl. (2023) 1–21.

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

[36] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, Houqiang Li, Semantic image synthesis via diffusion models, 2022, arXiv preprint arXiv:2207.00050.

[37] Gwanghyun Kim, Taesung Kwon, Jong Chul Ye, Diffusionclip: Text-guided diffusion models for robust image manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2426–2435.

[38] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, Mark Chen, Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021, arXiv preprint arXiv:2112.10741.

[39] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, Xinchao Wang, Metaformer baselines for vision, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[40] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[42] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[43] Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, Brent J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, J. Neural Eng. 15 (5) (2018) 056013.

[44] Yitong Li, Kafui Dzirasa, Lawrence Carin, David E. Carlson, et al., Targeting EEG/LFP synchrony with neural nets, Adv. Neural Inf. Process. Syst. 30 (2017).

[45] Liangyan Mo, Yuhan Wang, Wenhui Zhou, Xingfa Shen, Wanzeng Kong, A Bi-LSTM based network with attention mechanism for EEG visual classification, in: 2021 IEEE International Conference on Unmanned Systems, ICUS, IEEE, 2021, pp. 858–863.

[46] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, Mubarak Shah, Decoding brain representations by multimodal learning of neural activity and visual features, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2020) 3833–3849.

[47] Artem Obukhov, Mikhail Krasnyanskiy, Quality assessment method for GAN based on modified metrics inception score and fréchet inception distance, in: Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4, Springer, 2020, pp. 102–114.

[48] Yanfang Long, Wanzeng Kong, Xuanyu Jin, Jili Shang, Can Yang, Visualizing emotional states: A method based on human brain activity, in: Human Brain and Artificial Intelligence: First International Workshop, HBAI 2019, Held in Conjunction with IJCAI 2019, Macao, China, August 12, 2019, Revised Selected Papers 1, Springer, 2019, pp. 248–258.

[49] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, Mubarak Shah, Brain2image: Converting brain signals into images, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1809–1817.

[50] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Mubarak Shah, Generative adversarial networks conditioned by brain signals, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3410–3418.

[51] Ahmed Fares, Sheng-hua Zhong, Jianmin Jiang, Brain-media: A dual conditioned and lateralization supported GAN (DCLS-GAN) towards visualization of image-evoked brain activities, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1764–1772.